Deep Multi-Model Fusion for Human Activity Recognition Using Evolutionary Algorithms

Article *in* International Journal of Interactive Multimedia and Artificial Intelligence · December 2021

DOI: 10.9781/ijimai.2021.08.008

Kamal Kant Verma
University of Engineering and Technology Roorkee

Deep Multi-Model Fusion for Human Activity Recognition Using Evolutionary Algorithms

Kamal Kant Verma[1]*, Brij Mohan Singh[2]

[1] Research Scholar, Department of CSE, Uttarakhand Technical University, Dehradun (India)
[2] Department of Computer Science and Engineering College of Engineering Roorkee, Roorkee (India)

Received 30 October 2020 | Accepted 14 May 2021 | Published 6 August 2021

unir
LA UNIVERSIDAD
EN INTERNET

Abstract

Machine recognition of the human activities is an active research area in computer vision. In previous study, either one or two types of modalities have been used to handle this task. However, the grouping of maximum information improves the recognition accuracy of human activities. Therefore, this paper proposes an automatic human activity recognition system through deep fusion of multi-streams along with decision-level score optimization using evolutionary algorithms on RGB, depth maps and 3d skeleton joint information. Our proposed approach works in three phases, 1) space-time activity learning using two 3D Convolutional Neural Network (3DCNN) and a Long Sort Term Memory (LSTM) network from RGB, Depth and skeleton joint positions 2) Training of SVM using the activities learned from previous phase for each model and score generation using trained SVM 3) Score fusion and optimization using two Evolutionary algorithm such as Genetic algorithm (GA) and Particle Swarm Optimization (PSO) algorithm. The proposed approach is validated on two 3D challenging datasets, MSRDailyActivity3D and UTKinectAction3D. Experiments on these two datasets achieved 85.94% and 96.5% accuracies, respectively. The experimental results show the usefulness of the proposed representation. Furthermore, the fusion of different modalities improves recognition accuracies rather than using one or two types of information and obtains the state-of-art results.

Keywords

Human Activity Recognition, Support Vector Machine, 3D-Convolutional Neural Network, LSTM, Deep Learning, Genetic Algorithm, Particle Swarm Optimization.

DOI: 10.9781/ijimai.2021.08.008

I. Introduction

Human activity recognition (HAR) is a demanding research topic from last two decades in computer vision. HAR's objective is to learn dynamically and automatically of human activities such as drinking, eating, clapping, walking, etc. Human activities are mainly associated with daily human activities, indoor and outdoor activities. HAR has many applications in multiple domains, such as human security applications, health-care sectors, virtual reality games, intelligent monitoring systems, and human-computer applications [1]-[2]. Moreover, activity recognition is a challenging task, due to the structural changes among the subjects, inter class and intra class similarity between the activities. Additionally, some issues still exist, such as cluttered background, occlusion, camera motion, multi-camera recognition, complex scenes, human to human interaction, or human to object interaction that makes video-based human activity recognition systems more complex and challenging [3]. Systematically analyzing and recognition of human postures can make people understand the sense of human behavior. This could be advantageous to augment the monitoring process of indoor activities, understand normal and abnormal activities, and to maintain security surveillance systems in real life. Hence, the human activities recognition system has become a vital issue nowadays in both academia and industry.

* Corresponding author.
E-mail address: kkv.verma@gmail.com

From the past few years, there is a significant improvement in this research area, due to two main factors. Firstly, frequent accessibility of low-cost depth camera, secondly the robust features learning using deep convolutional neural network. Before the advent of the RGB-D (Depth sensor), the research on activity recognition was limited to the recognition of human activities from frame sequences captured using traditional RGB video cameras [3]. As the depth camera like Microsoft Kinect, Xbox 360 Kinect, and ASUS Xtion came into existence, the research has shifted towards the learning of human activities using depth information [4]-[6].

Unlike the RGB, the depth camera has several advantages, for example, there is no effect of variable lighting conditions, illumination changes. Segmentation of shape and structure given by the depth camera are easy to handle, insensitive to the cluttered background. It also gives additional modality like 3D skeleton joint positions that deliver multi-dimensional data for better activity recognition. Zhu et al. [7] proposed a robust feature-fusion approach using 3D-skeleton joints information to capture spatial-temporal features.

Currently, human activity recognition techniques can be classified into two groups: traditionally handcrafted feature extraction techniques and using deep learning techniques. In the previous study, the traditional handcrafted features extraction methods include Histogram of Oriented Gradients (HOG) and Histogram of optical flow (HOF) from 2D-images [9]. In a recently published article, Franco et al. [10] used RGB video sequences and skeleton joint positions for activity recognition. They used HOG to learn motion information

Please cite this article in press as:
K. K. Verma, B. M. Singh. Deep Multi-Model Fusion for Human Activity Recognition Using Evolutionary Algorithms, International Journal of Interactive Multimedia and Artificial Intelligence, (2021), http://dx.doi.org/10.9781/ijimai.2021.08.008

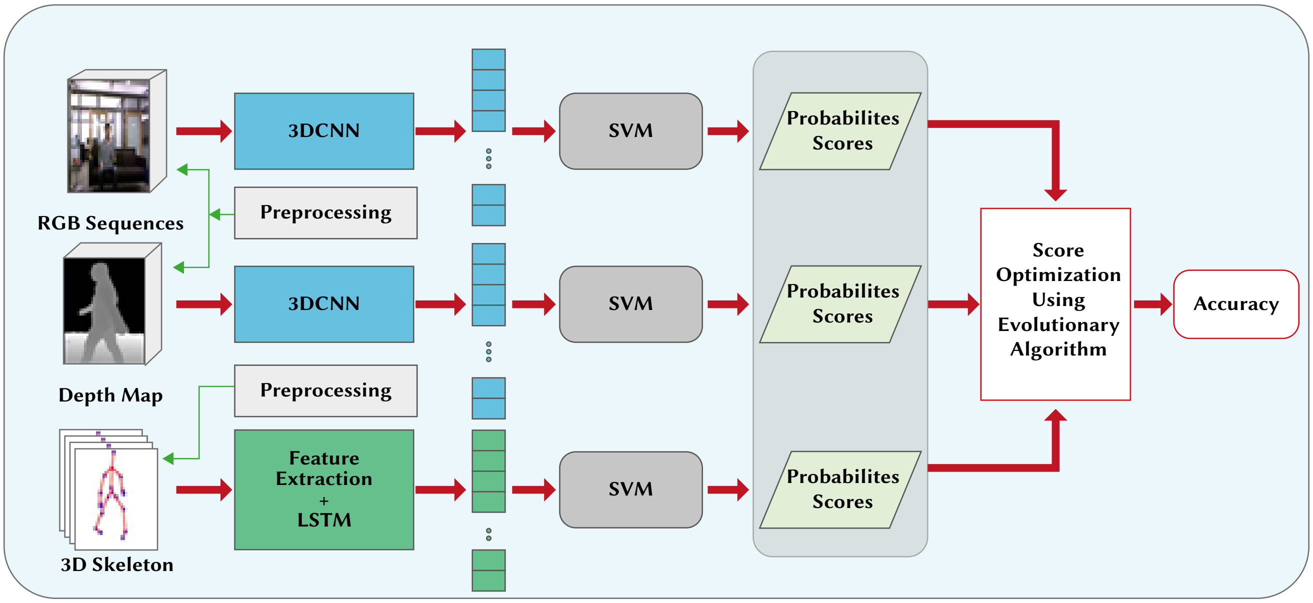


Fig. 1. Flow diagram of the proposed system.

from RGB video sequences and bag-of-words (BOW) for spatial information learning from the skeleton postures. However, the skeleton joints position information extracted from the depth map contains several difficulties like occlusion due to any arbitrary object present in the scene, lack of precision, and viewpoint variation [11]. Therefore, to prepare a robust action recognition system, these errors can be eliminated by finding the body pose estimation component. Hence, Chaaraoui et al. [8] proposed a method that combines body pose estimation and 2D shape clue parameters to make a robust action recognition system. They extracted the low dimensional features from skeleton joints, fetched silhouette from depth images, and combined both to improve the activity recognition. Pham et al. [12] and Yang et al. [13] also used in-depth information about the 3D structure of human body parts and body posture information.

Similarly, Jalal et al. [14] also used human silhouette as additional information from depth video. They used R-transformation to find the translational and scaling view-invariant features, the principal component analysis (PCA) is applied to make the features low dimensional. Furthermore, they also used linear discriminant analysis (LDA) to obtain prominent features, and finally, the Hidden Markov Model (HMM) is used to train and identify human activities.

However, the methods discussed in the above section only utilize handcrafted features. Though the extractions of the handcrafted features are a challenging task, and they do not require a large amount of training data. Moreover, it is a time-consuming process also, and these are the standard methods and are not ideal for a specific task such as activity recognition from complex scenes. Therefore handcrafted features may not thoroughly learn the intrinsic spatial and temporal information present in RGB-D sequences thus, these methods could not satisfy as per desire.

From the past few years, deep neural networks have gained considerable achievements in computer vision using deep architectures. A deep neural network is an ordered arrangement of multiple layers that learns the features automatically. In DNN, each layer gets the output as an input from the previous layer and performs nonlinear transformation. One such deep neural network is convolutional neural network (CNN) that gets better recognition results. In the era of computer vision, CNN achieves enough popularity in vision based applications such as surveillance [15], image classification [16], image segmentation [17], object classification, medical sector [18], object recognition [19], recognition of facial expression [20], biometric and agriculture [21], video classification [22] and many others [23]. CNN is enough capable to handle a large amount of data with a high level of accuracy [18].

However, the trainable feature extraction using CNN called CNN-features and recognition of CNN-features using SVM (CNN-features + SVM) architecture has achieved remarkable advantage in the previous literature. For example, Niu et al. [24] evaluated CNN-features + SVM architecture on 10000 grayscale images of MNIST dataset, and showed that the CNN-features + SVM combination generated highest 99.81% recognition accuracy. Similarly, Xue et al. [25] have proposed NBI-Net architecture for cancer detection which is again based on CNN as trainable feature extractor and SVM as predictor. The proposed architecture was trained and tested on 6500 patch samples and achieved 90.93% recognition accuracy which was 0.9% more as compare if only CNN had been used alone for recognition. Likewise, Sargano et al. [26] have also used CNN-SVM architecture for human activity recognition. They used pre-trained CNN network as trainable feature extractor and SVM for recognition of human activities. The proposed architecture has been evaluated on two UCF-Sports and KTH datasets and achieved highest accuracy over mentioned state-of-the-art methods. Inspired by the above discussed outstanding performance we have also used CNN-features + SVM architecture to process RGB and depth data. Another reason for considering CNN-SVM architecture is because both CNN and SVM have already shown excellent performance on human activity recognition independently. Therefore, in this work we have focused on their fusion to come out their best qualities. On the other hand, RNNs networks have also shown the exceptional performance on sequence modeling applications such as image captioning [27], video analysis [28], and language translation [29]. Due to this reason, this work used RNN especially LSTM network to process skeleton data.

Likewise, deep multi-model approaches [30]-[32] have substantially outperformed over the manual features extraction methods [11]-[14]. Therefore, this paper also suggests a multi-model approach which uses CNN-features + SVM architecture to process RGB, depth data and a LSTM network to process skeleton joint positions. The SVM classifier with RBF kernel has been used corresponding to classification using features extracted in each parallel channels for RGB, depth and skeleton data. The flow diagram of our proposed approach is given in Fig. 1. The proposed approach uses three different types of neural network: two 3DCNN for RGB and depth videos and a LSTM network for skeleton stream.

- First, spatio-temporal feature learning has been implemented from RGB and Depth video sequences using two different 3DCNN models. At the same time, a LSTM network has been trained using a feature vector obtained from the skeleton data.
- Secondly, in order to find the class score of each test activity from individual streams, three SVM networks have been trained using the spatial-temporal features that have been extracted using three different deep neural networks for each modality.
- At last, an evolutionary algorithm is used to optimize the class score obtained for the three SVM networks during score-level fusion from the individual streams to find the optimized class label of each test activity.
- We have trained and tested our proposed method on MSRDailyActivity3D and UTKinectAction3D Datasets.

The remaining sections of the paper are organized as: The relevant literature review is given in the section II. Section III contains our proposed work for activity recognition. Section IV consists of experimental work, results and discussions. Section V contains conclusion followed by future work.

II. Related Work

Activity recognition using RGB-D is a hot research area from past several years. Hence, in this section we covered previous literature using RGB, Depth videos, skeleton joints and hybrid data.

A. Activity Recognition Using RGB Data Only

Activity recognition in RGB videos is a most frequent task because of availability of RGB video activity datasets, for example Soomro et al. [33] developed the UCF-101 dataset containing 101 different activity classes with more than 13K video clips. Kuehne et al. [34] developed another commonly used dataset HMDB-51 containing 51 activity categories with 7000 video clips. In addition to that, Caba Heilbron et al. [35] proposed the large scale activity dataset ActivityNet. The ActivityNet dataset contains 203 different activity classes with approximately 137 untrimmed video clips per class. Besides these datasets, several others RGB datasets also exist for activity recognition task [36]. Prior to the learning based approaches, numerous handcrafted-based feature descriptors methods have been suggested by the researchers in past literature [36]. A lot of handcrafted-features based approaches are presented in [36], that are ranges from space-time features descriptors, appearance-based (shape and motion) and fuzzy logic based approaches, etc. Out of them dense trajectory [37] methods have achieved best results compare to others.

Later, the convolutional neural network has been efficiently used for activity recognition in RGB video classification due to the remarkable performance of deep neural network (DNN) in image and video classification [38]. Simonyan et al. [39] proposed a two-stream action recognition approach for spatial and temporal information. The spatial information is extracted from the frame using spatial stream convolutional neural network (ConvNet) and temporal information is extracted from the frames using optical flow displacement field vectors (OFDF). Both the streams are combined together for late fusion. The average fusion and multi-labels SVM are used for classification. The method is tested on UCF-101 and HMDB-51 datasets and obtained 88.0% and 59.4% accuracy respectively. Karpathy et al. [40] proposed multi-resolution CNN with two streams. The first stream is fovea stream and second stream is context stream. Both the streams take the consecutive frames as an input. The extracted information from both the streams are fused using different fusion techniques such as early, late and slow fusion. The proposed approach is validated on large scale activity dataset named sports 1M that contains over one millions sports videos from youtube videos which is divided into 487 different classes. Tran et al. [41] used 3DCNN to learn spatio-temporal features from the video sequences simultaneously. They applied varying kernel size and found that 3×3×3 kernel outperforms among the used kernel size. The proposed approach in [41] produces superior results on four datasets: UCF-101, ASLAN, YUPENN and UMD datasets. Feichtenhofer et al. [42] suggested the features-level fusion strategy to take the benefit of spatio-temporal information. Spatial-temporal features fusion at the last convolution layer are more advantageous rather than fusing at softmax layer and also decreases the parameter counts. The proposed approach is implemented on two benchmark datasets UCF-101 and HMDB-51 and gave 93.5% and 69.2% accuracies. Varol et al. [43] have given a convolutional neural network based on long-term temporal information (LTC-CNN). They implemented LTC-CNN with varying number of temporal information ranges from t=16 to t=100 and found that as the temporal information t increases, the value of the accuracy also increases. They experimented their approach on two challenging datasets, UCF-101 and HMDB-51, and achieved 92.7% and 67.2% accuracies respectively. Ullah et al. [44] gave a novel framework for activity recognition using CNN and deep-BLSTM. In this framework deep features are extracted from the every sixth frame of a video sequence using CNN. Next, the features learning have been performed by deep-BLSTM network. This approach is tested on UCF-101, Youtube-11 and HMDB-51 datasets and obtained 91.21%, 92.84% and 87.64 accuracies comparable to the state-of-the-art results. Recently Verma et al. [45] used powerful coarse and file level classification framework for single and multi-limb activity recognition. The advantage of this framework is that, it divides the losses at multiple levels. They implemented their approach on UTKinectAction3D dataset and achieved 99.92% accuracy at coarse level with overall 97.88% accuracy.

For the spatial-temporal activities learning from the RGB information, most of the above discussed methods used deep convolutional neural networks. Unlike the above methods, the proposed approach uses deep convolution neural networks for feature learning with support vector machines for score prediction from RGB streams.

B. Activity Recognition Using Depth Data Only

With the advent of depth cameras such as Kinect, Xbox 360, and ASUS Xtion, depth video sequence drastically changed the research pattern. It has several advantages such as it is constant with respect to the background changes and contains depth information which is not present in RGB. Recently, Li et al. [46] proposed an idea of mapping the frames onto three orthogonal planes top view, side view and front view, then they formed the depth motion map (DMM) by stacking all three images. Local ternary pattern are applied to DMM to filter out the images and finally they used CNN for classification. The method in [46] gained state-of-the-art accuracy on MSRAction3D and MSRGesture3D datasets and achieved 98.81% and 99.67% accuracies respectively. Wang et al. [47] proposed Dynamic Depth Image (DDI), Dynamic Depth normal image (DDNI) and Dynamic Depth Motion Normal Image (DDMNI) form the depth videos. DDI is used to capture motion dynamics of the sequence while the structural information of a posture is extracted using both DDNI and DDMNI. At last, convolutional neural network (ConvNet) is used for activity classification. The proposed method is implemented on Large-scale Continuous Gesture Recognition, Large-scale Isolated Gesture Recognition and NTU RGB+D datasets, and achieved 59.21%, 87.08% and 84.22% cross view accuracies which is greater than state-of-the-art accuracies. In addition to that, Chen et al. [48] also presented an effective depth motion map based local binary pattern (DMMs-LBP) framework for activity recognition. In this approach three motion maps have been generated using frame differencing after mapping the depth images onto three orthogonal planes related to top view, side view and front view. Then, a local binary pattern (LBP) descriptor has been used to find the LBP histogram for each DMM and represented by a feature vector. The method suggested in [48] is implemented on MSRAction3D and MSRGesture3D datasets and obtained 87.9% and 96.4% accuracies respectively. Similarly, Wang et al. [49] also proposed a framework known as hierarchal depth motion map (HDMM) with three channel convolutional neural networks related to front, side and top view. The method is tested on MSRAction3D, MSRAction3DExt, UTKinectAction3D and MSRDailyActivity3D datasets and achieved state-of-the-art results. However, Megavannan et al. [50] used handcrafted features prior to the popularity of deep neural network. They used motion history images (MHI), average depth image (ADI) and depth difference images (DDI) in order to find the space-time features. Then two features have been generated, firstly Hu movement feature is calculated using MHI and ADI and in the second case DDI is used to find the hierarchically division. They also created their own dataset having eight activities to validate their approach and obtained overall 90% accuracy.

Even though most of the above methods used depth motion map (DMM) to learn temporal information, our spatial-temporal feature learning policy for depth channel utilizes 3D deep convolutional neural network (3DCNN) and further find the class score using conventional support vector machine. This combination of 3DCNN+SVM outperforms over traditional handcrafted features and DMM strategy.

C. Activity Recognition Using Skeleton Data Only

Besides the limitation of the depth maps, the skeleton joint positions have an advantage that it is three dimensional in nature compared to the one dimensional depth map and two dimensional RGB data. Furthermore, skeleton positions are also considered as temporal information, therefore many past approaches have attempted to use RNN network specially LSTM. Recently, Han et al. [51] proposed global spatio-temporal (GL-LSTM) model for activity recognition. The GL-LSTM architecture combines the accumulative learning curve (ALC) for temporal information and global spatial attention (GSA) to prepare the spatial information. Then LSTM network with difference clue has been used for action classification. This approach is validated on NTU RGB-D and SBU datasets and generated results outperform the given state-of-the-art methods. Similarly, Ren et al. [52] also discussed several action recognition methods using deep recurrent neural network (DRNN), convolutional neural network (CNN), and graph convolutional neural network (GCNN). They also discussed latest skeleton datasets with their performance. Another skeleton based coarse and fine level continuous human activity recognition framework is suggested by Saini et al. [53]. In this, all activities have been segmented into two categories such as sitting and standing. Next, five features such as 3D-joints, angular movement, angular direction, distance and velocity have been extracted for activity classification using BLSTM network. For evaluation of the approach, they captured 1110 continuous activity sequences for 24 activity classes using Kinect depth sensors and achieved 68.9% and 64.45 % accuracies through length modeling and without length modeling respectively. Chikhaoui et al. [54] proposed a 3D skeleton based aggressive behavior learning framework, this framework is based on the fusion of two features such as joint based and body part based in order to learn the spatio-temporal information. Then the combined feature vector is used by ensemble learning based rotation forest. The proposed approach is tested on various 3D activity datasets for instance TRI, Kintense, UTKinectAction3D, Florence Action and MSRAction3D datasets and absolutely distinguishes the various activity categories for each dataset. Shahroudy et al. [55] suggested a novel part aware LSTM (P-LSTM) in which local structure of five different body parts (two legs, two hands and torso) are individually mapped. These five different memory cells are combined to find the global information. They have evaluated their work on a self-developed NTURGB-D dataset and have shown the effectiveness of the proposed P-LSTM network on the traditional recurrent neural network.

The above discussion exhibits the better outcome of DRNN specially LSTM for activity recognition using 3D skeleton joint positions. It also shows that the LSTM network has improved performance over other deep recurrent neural networks for sequence learning problem in 3D domain. Therefore, we have also used LSTM network to learn space-time features from skeleton data.

D. Action Recognition Using Hybrid (RGB+Depth+Skeleton) Data

In past literature, combining the RGB, depth and skeleton joints information has given promising results for human activity recognition. Recently, Gu et al. [56] proposed a novel framework which combines domain knowledge clue parameter for decision making. Three motion history images (MHIs) have been used to learn global information. Local spatial and temporal information are extracted using skeleton joint positions, and this information is fused together with domain knowledge clue parameter. The proposed work is evaluated on two RGB-D datasets and has given the best results on the mentioned stated-of-the-art methods. Khaire et al. [57] proposed a 5-stream CNN for activity recognition using MHI, DMM and skeleton images. Moreover, this work has given a new way of creating skeleton images. The given approach is implemented on CAD-60, SUB Kinect interaction and UTD-MHAD datasets and got comparative results on the state-of-the-arts methods. Tomas et al. [58] presented a method that uses CNN and stacked Auto Encoder (SAE). CNN is used to learn motion information from motion history images and SAE is used to find the physical structural information of a static posture. The proposed method is tested on two benchmark datasets MSRDailyActivity3D and MSRAction3D and achieved 91.3% and 74.6% accuracy respectively. Ijjina et al. [59] proposed an approach that utilizes the key pose related to each action. They captured motion information using two temporal MEI and MHI. Then they took the frame differencing from each category. The temporal information using MEI and MHI have been captured from RGB and depth video sequences and placed as an input to the CNN for action recognition. The suggested method is implemented on four RGB-D datasets namely, MIVIA action, Weizmann, SBU Kinect action, and NATOPS gesture datasets and achieved 93.37%, 100%, 90.98% and 86.58% accuracies. Zhao et al. [60] also proposed a multi-model approach for human behavior recognition using RGB, depth and skeleton data. They computed space-time features from each modality using deep 3DCNN and obtain class probability scores of each test activity using SVM. Next, achieved class probability scores of each test activity are fused together using weighted linear combination techniques. Subsequently, they implemented their approach on two RGB-D datasets such as MSRAction3D and MSRDailyActivity3D, and achieved 94.15% and 97.29% accuracies.

After the above discussion, some of the approaches use all three modalities like RGB, depth and skeleton joints positions, while some use RGB and skeleton, depth and skeleton while some others use RGB and depth sequences. Therefore, next, in the section III, we have discussed our Deep Multi-Model (DMM) fusion approach which combines RGB, depth sequence and skeleton joints positions for human activity recognition system.

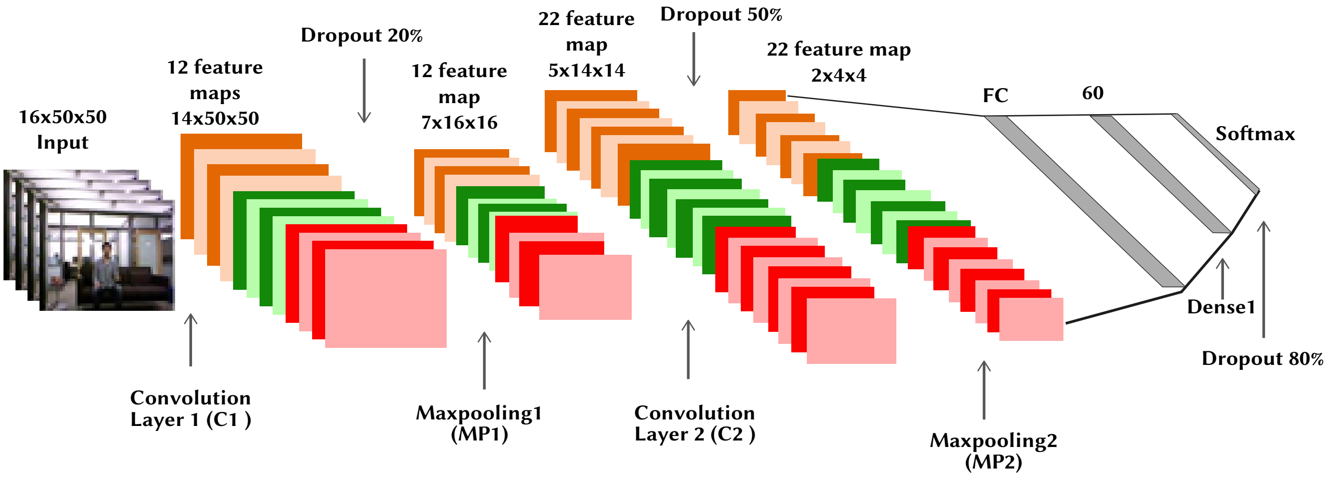


Fig. 2. Architecture of 3DCNN for RGB video sequences.

III. Proposed Work

In the proposed work to recognize the human activity, a novel multi-model approach with different modalities –RGB, Depth and skeleton joint information –using evolutionary algorithms has been given. The proposed work is done in three different levels. In the first level, spatial-temporal features have been extracted from different modalities, in the second level three independent SVMs are trained using the features extracted from the first level, and in the last level the probability scores are fused and optimized using two evolutionary algorithms such as GA and PSO.

The spatial-temporal activity learning from each modality has been discussed in the subsequent sections.

A. Spatio-Temporal Activity Learning From RGB Information Using 3DCNN

The 3D Convolutional neural network (3DCNN) is a deep neural network introduced in [61], used to learn the features from both spatial and temporal dimension. The 3D convolutional is accomplished by convolving a 3-dimensioanl filter over the cube obtained by stacking the spatial and temporal frames one after another. In order to learn the motion related information from the sequence of frames, features maps exist in the convolution layer that are connected with the multiple contiguous frames from the previous layer. Multiple convolutional layers are used to obtain both lower and higher level features. Hence the design technique of Convolution neural network is the increase the feature maps by increasing the number of layers in the network. Therefore, a 3D Convolutional neural network is obtained by convolving the 3D filter kernel over the multiple staked fame to produce a 3D cube. Finally, the value in the jth feature map of ith layer at a position (x, y, z) is given in equation (1).

$$v_{i,j}^{x,y,z} = \tanh\left(b_{ij} + \sum_m \sum_{a=0}^{A_{i-1}} \sum_{b=0}^{B_{i-1}} \sum_{c=0}^{C_{i-1}} w_{ijm}^{abc}\, v_{(i-1)m}^{(x+a)(y+b)(z+c)}\right) \quad (1)$$

Where m is the index value of the feature maps in the (i-1)th layer, which are connected to the present feature map and b_{ij} is the bias value of the mth feature map. The function tanh () is the hyperbolic tangent function. The values A_i and, B_i represent the height and width of the 3D-kernel and C_i is the dimension of the 3D-kernel along the temporal direction. The term w_{ijm}^{abc} is the (a, b, c)th value of the kernel of the mth feature map in the previous layer.

In order to learn the spatio-temporal features from the RGB video sequences, a 3DCNN is used. The proposed 3DCNN contains two convolution layers and two max-pooling layers. To prepare the input for the 3DCNN network, simple preprocessing has been performed in which each frame is resized to a fixed height and width dimension. The height and width is resized to [50, 50] using the Open CV library in python. Since each video sequence has a varying number of frame length so it is difficult to normalize all video sequences to the same length, hence a fixed input sequence is used as a depth dimension. The value of the input depth dimension was set to 16. Therefore the size of the RGB input cube which is given as an input to the 3DCNN is [16(depth) ×50(height) ×50(width)]. This input cube is processed by the first convolution layer (C1) followed by the first maxpooling layer (ML1). A dropout layer with an amount of 20% is used after the first convolution layer. The network also contains a second convolution layer (C2) followed by a second maxpooling layer (MP2). A 50% dropout is also used between the C2 and MP2. Then a fully connected layer (FC) layer is used to obtain the one dimensional feature vector. Then, a dense layer with 60 neurons is used. Finally a softmax layer is used for classification followed by a dropout layer with 80% dropout amount. The number of feature maps in C1 and C2 are 12 and 22 respectively. The size of the kernel in first and second convolution layers are (3×1×1) and (3×3×3) respectively. For the first and second 3D maxpooling layers, (2×3×3) down sampling is used for both MP1 and MP2 layers. We have used the same network parameter to process RGB video sequences for both MSRDailyActivity3D and UTKinectAction3D datasets. The architecture of the 3DCNN for spatio-temporal features learning from RGB video sequences is shown in Fig. 2.

B. Spatio-Temporal Activity Learning From Depth Information Using 3DCNN

Similarly, for the spatio-temporal feature learning from the depth map sequence, a different 3DCNN is used with different network parameters. During the preprocessing step, the input cube is formed in the similar way as we performed in RGB, except that different length, width and depth dimension are applied. The size of the input depth cube which is inserted to the 3DCNN is [13(width) ×32(height) ×32(width)]. The input depth cube is processed by first 3D convolution layer (DC1) followed by maxpooling layer (DMP1). A dropout of 20% has been used in between. Another 3D convolutional layer (DC1) is used followed by a second (DMP2) layer. Similarly, 20% dropout is also used between the second 3D convolutional layer and second maxpooling layer DC1 and DMP2. Then, a fully connected (FC) layer is used to find the one-dimensional feature vector, followed by a dense layer with 128 neuron units. At last a softmax layer is used for classification. We have also used a dropout layer with 50% amount before the softmax layer. The count of feature maps for DC1 and DC2 layers are 16 and 32. The size of the kernel used in DC1 and DC2 layers are (3×3×3) and (5×5×5) respectively. The downsampling size of 3D

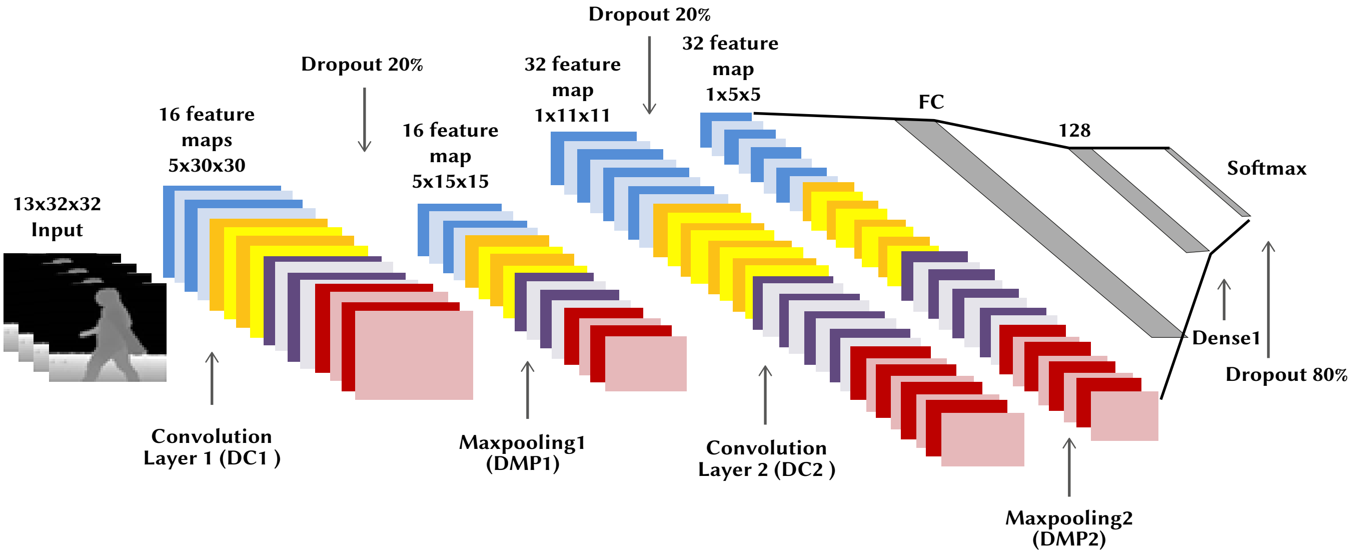
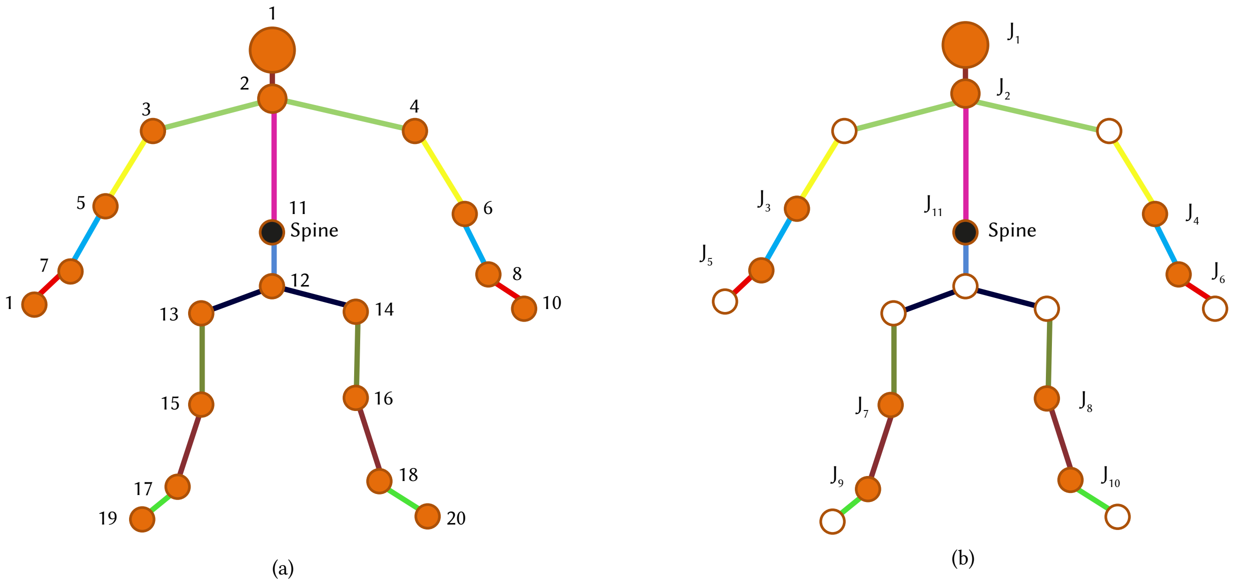


Fig. 3. The architecture of 3DCNN for spatio-temporal feature learning from depth maps.



(a)

(b)

Fig. 4. 3D joints selected as per the evolutionary algorithm where J represents the positions of 11 body joints.

maxpooling layers are (2×2×2) and (1×2×2) respectively. We have used the same network settings to learn the spatio-temporal feature from the depth maps for both MSRDailyActivity3D and UTKinectAction3D datasets. The architecture of 3DCNN for spatio-temporal features learning from depth videos is given in Fig. 3.

C. Spatio-Temporal Activity Learning From Skeleton Joints Information Using LSTM

In the present section, we have extracted the set of relevant and observable features from the skeleton joint sequence and prepared a feature vector (F_T). Three features corresponding to position, spatial and temporal dimension are used to make a feature vector (F_T). In order to make F_T we used, 3D joints position (F_{3DJ}), Minkowski distance (F_{MD}) and Temporal (F_{Temp}) features. The feature vector is given in (2)

$$F_T = \{F_{3DJ}, F_{MD}, F_{Temp}\} \qquad (2)$$

1. 3D Joints Positions Feature

A twenty 3D body joints corresponding to an activity captured by a kinect sensor in the dataset are shown in Fig. 4(a). However, not all joints are informative and useful for activity recognition due the unwanted noise present in the device during data capturing and the orientation of the human body. Therefore, a minimal set of joints positions are recognized using an evolutionary algorithm which determines the optimal set of 3D joints positions. For this purpose, we have used the method performed in [63] that eliminates the redundant joints positions because of closeness between the joints, and provides the optimal set of joints positions for activity recognition. For example, joints (wrist and ankles) are redundant to joints (hands and feet) and not informative for activity recognition. Using the above concept, a feature (F3DJ) is extracted having eleven 3D joints positions as shown in Fig. 4(b). The optimal set of joints position is given in (3)

$$F_{3DJ} = \{J_1, J_2, J_3\, J_4 J_5 J_6, J_7, J_8, J_9, J_{10}, J_{11}\} \qquad (3)$$

Where $J_1, J_2, J_3 \ldots\ldots J_{11}$ are the positions of the optimal body joints.

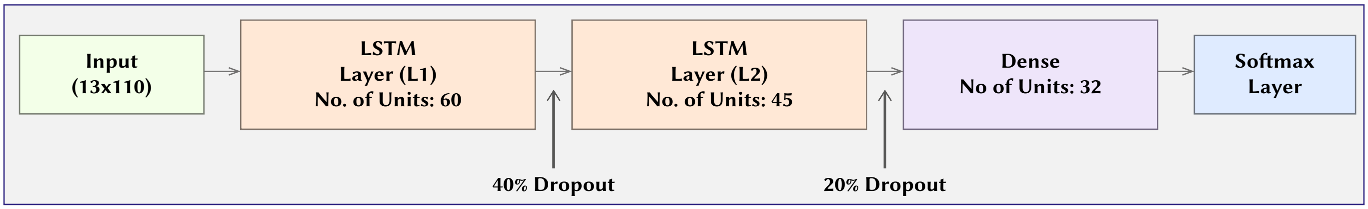


Fig. 5. Architecture of LSTM for spatio-temporal features learning from Skeletons joints Sequences.

2. Minkowski Distance Feature

To learn the spatial aspects of an activity sequence, we have utilized the concept of symmetric matrices using the 3D body joints positions. The Minkowski distance is the generalization of the both Euclidian and Manhattan distances.

Here, the Minkowski distance feature has been calculated from the optimal joints say N obtained from the equation (3). Here the value of N is 11 from the equation (3). The Minkowski distance between the two points X = (x_1, x_2, $x_{3,}$.......x_n) and Y = (y_1, y_2, $y_{3,}$.......y_n) of order r is defined in (4).

$$Dist_r(X,Y) = (\sum_{I=1}^{n} \lceil x_i - y_i \rceil^r)^{1/r} \quad (4)$$

Finally, a 55-dimensioal feature vector is generated corresponding to the Minkowski distance matrix.

3. Temporal Feature

To learn the temporal information from the activity sequence of 3d joints we first find the maximum and minimum values of the coordinate of any joint J_i, then for each frame in the sequence we obtain the difference between the coordinates for the frame to the maximum and minimum values of the same 3D joint for a complete sequence. For a 3D joint J_i, and 3D coordinates system (J_{ix}, J_{iy}, J_{iz}), , we obtained($J_{i,max}$, $J_{i,min}$) as given in the equation (5) and (6) respectively.

$$J_{i,max} = \frac{(\max(J_{ix_t}) - J_{ix}) + (\max(J_{iy_t}) - J_{iy}) + (\max(J_{iz_t}) - J_{iz})}{3} \quad (5)$$

$$J_{i,min} = \frac{(J_{ix} - \min(J_{ix_t})) + (J_{iy} - \min(J_{iy_t})) + (J_{iz} - \min(J_{iz_t}))}{3} \quad (6)$$

Where ⟨max (J_{iy_t}), max (J_{iy_t}), max (J_{iz_t})⟩ and ⟨min (J_{iy_t}), min (J_{iy_t}), min (J_{iz_t})⟩ are the maximum and minimum values of the coordinates of joint J_i for all the sequence respectively. Finally, a 22 dimensional temporal feature (F_{Temp}) vector is obtained by combining all the minimum and maximum values such as

F_{TEMP} = {$J_{1,min}$, $J_{2, min}$, ...$J_{N, min}$, $J_{1,max}$, $J_{2, max}$, ...$J_{N, max}$}

4. Feature Learning Using LSTM

The LSTM network is proposed by Hochreiter and Schmidhuber [62] in 1996. Not at all like other RNN, LSTM protects and keeps up the errors so that they can be effortlessly backpropagated through layers, which also makes LSTMs valuable for time series predictions and makes the model to continuously learn using a wide number of time steps. A LSTM network contains input gate, output gate, forget gate and memory units in the recurrent layer. The memory units contain the memory cells which are used to maintain the temporal state of the network using self-connections. These gates performed the function according to the received signal and pass or block the information based on its strength. The gate weights are used to filter these signals. These weights act in the same way as ordinary NN input and hidden states as they are learned all through the RNN learning prepare.

The input gate is utilized to assure that the all of data included to the cell stated is vital and not redundant and it controls the flow of input into the memory cells. The careful selection of valuable data from the current cell state and showing it as an output is done with the output gate. A forget gate is outlined for annihilating information from the current cell state. The LSTM execution is optimized by the evacuation of any meaningless information that is not required by the LSTM to get it things or to evacuate any information that is not important anymore. An LSTM network performs a mapping from the input sequence X = {x_1, x_2, x_3..... x_t} to the output sequence Y = {y_1, y_2, y_3..... y_t} using the network activations reclusively from t = 1 to T.

The space time feature vector (F_T) given in (2) is used to train the LSTM network, which contains two LSTM layers with 60 neurons and 45 neurons respectively. Two dropout layers are used after each LSTM layer with 20 % dropout in each case. Then a dense layer is used with 32 neurons. Finally a softmax layer is used for output classifications. The rectified linear unit 'ReLU' activation function is used in each LSTM layer. The architecture of LSTM is given in Fig. 5.

D. Score Prediction Using SVM

The SVM classifier is based on the kernels [64]. The main objective of this classifier is to map the input samples into the higher dimensional feature space, and insert a hyperplane that distinctly classify the input samples. SVM performs both types of classification, linearly and non-linearly, with the help of different kernels. Some of the kernels of SVM are linear, radial bases function (RBF), polynomial, etc. In our work, one-vs-all SVM classifier with radial basis function (rbf) kernel is trained using the training features extracted from the second last layer (layer before softmax) of the network for each models independently. To make one-vs-all SVM classifier, the value of decision_function_shape parameter has been set to 'ovr'. During the testing, the corresponding test features have been extracted for each test sequence and applied to the trained SVM model. Then, the classification probability scores of each test sequence are calculated. Finally, the obtained probability scores for each test sequence of the three models are fused together for optimization using evolutionary algorithm.

E. Score Optimization Using Evolutionary Algorithm

In this section, we used two optimization algorithms, GA and PSO separately to find the optimized class score of each test activity.

1. Score Optimization Using Genetic Algorithm

Genetic Algorithm is developed by Goldberg [65] in 1989. GA is a heuristic search method to solve both unconstrained and constrained optimization problem which is based on the process of natural selection. GA iteratively changes the population of individual solutions. The algorithm picks individuals from the current population in each step and utilizes them as a parent to create the population for the next generation. The population generates the optimal solution after the continuous iterations. The process of the genetic algorithm starts with an initial population and the choice of the initial population generally depends on the optimization problem. The next process determines the selection of the proportion of existing population to evaluate the objective function. The ratio of the current population is taken to breed the new generation and fitness function is used to select the individual solution. Then the crossover and mutation process is used to generate a second level population of solutions. In

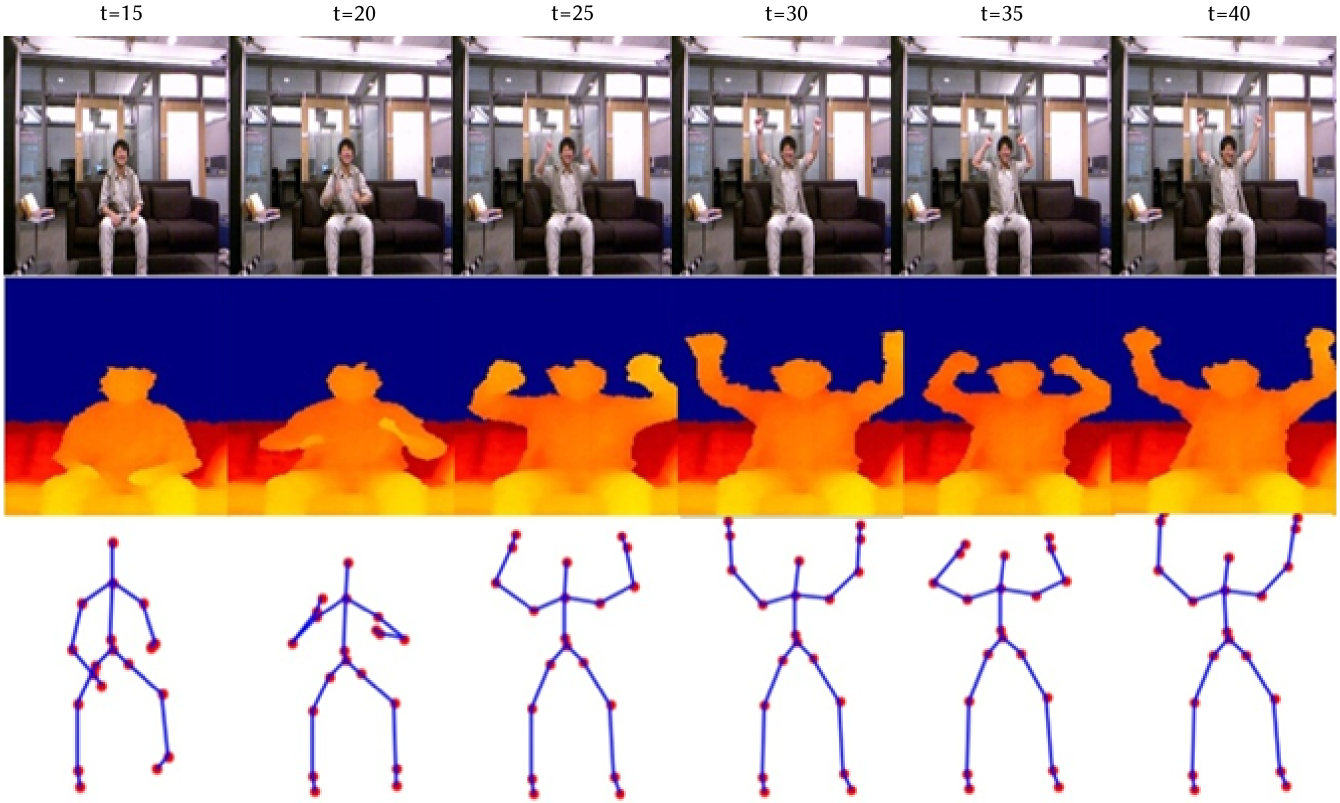


Fig. 6. Sample frames of 'cheer up' in MSRDailyActivity3D dataset at different time stamp in the three formats RGB, Depth and Skeleton.

the crossover process the selected parents are combined to generate the new population. The mutation process is used to produce the children by making the random change in parents. Similarly, the same process start again by inserting the new generation into the selection process and the algorithm repeats itself until some termination condition is satisfied. Some of the termination conditions are when the objective function reaches some threshold condition, fixed number of generations reached, or time limitation etc.

In this work, GA is used to optimize the probability scores for each activity sequences obtained from the trained SVM classifier for each modality. The optimal unbounded weights have been recorded against the highest accuracy. In addition to GA optimization, we have also used the PSO algorithm to optimize the probability score and compared the results of both optimizations.

2. Score Optimization Using PSO

The PSO is a well-known optimization algorithm based on population developed [66]. The PSO attempts to find the optimal solution of the problem using the population of particles. The basic principle of PSO is one in which each particle in the swarm represents an individual. And the combination of particles is a swarm. The PSO begins working in parallel, with a collection of particles, and reaches to the optimal solution using the current velocity, its prior best velocity and velocity of its neighbor particles. Some of the key features of the PSO are simple, effective, easy to implement and it does not required gradient information. The search space is considered as the solution space in PSO and a position in the search space signifies the solution of the problem. Different particles move in the search space with its velocity to find the optimal solution in the search space. The particle movement at each iteration is described in equation (7) and (8)

$$X_p(t+1) = X_p(t) + v_p(t) \tag{7}$$

$$v_p(t+1) = \omega v_p(t) + a_1 r_1 \left(xbest_p(t) - X_p(t)\right) + a_2 r_2 \left(ybest_p(t) - X_p(t)\right) \tag{8}$$

Where $X_p(t)$ is the position of the particle p at time t, $v_p(t)$ is the velocity of the particle at time t, $xbest_p(t)$ is the best position of the particle found by itself. $ybest_p(t)$ is the best position determined by all the swarm. ω is the inertia weight, a_1 a_2 are the two acceleration coefficients and r_1 r_2 are the two random variables whose value ranges between 0 and 1.

In this phase, the obtained probability scores are also optimized using PSO. The optimal weights corresponding to highest accuracy have been achieved after the successive iteration of the PSO algorithm.

IV. Experimental Results and Discussions

In this paper, we used the MSRDailyActivity3D and UTKinectAction3D datasets to train and test our proposed approach. Both datasets were captured using Kinect Senor. Activities in both datasets are synchronized in all formats such as RGB, Depth, and Skeleton. To train our model we used Intel Core i7 8th generation, 2.6 GHz processor with 16 GB of RAM, on Ubuntu 16.04 LTS (Linux) operating system having 2 GB 940MX NVIDIA GPU support. To implement CNN and RNN networks, we use the Keras Deep learning framework with version 2.2.4, and for implementing the Genetic algorithm optimization, we used MATLAB version 18a.

A. MSRDailyActivity3D Dataset

The MSRDailyActivity3D dataset [67] was captured with the help of Kinect sensor. It contains 16 human activities such as: 'drink', 'eat', 'read book', 'call cellphone', 'write on a paper', 'use laptop', 'use vacuum cleaner', 'cheer up', 'sit still', 'toss paper', 'play game', 'lie down on sofa', 'walk', 'play guitar', 'stand up', 'sit down'. The dataset is captured by ten subjects, out of them five are male and five are

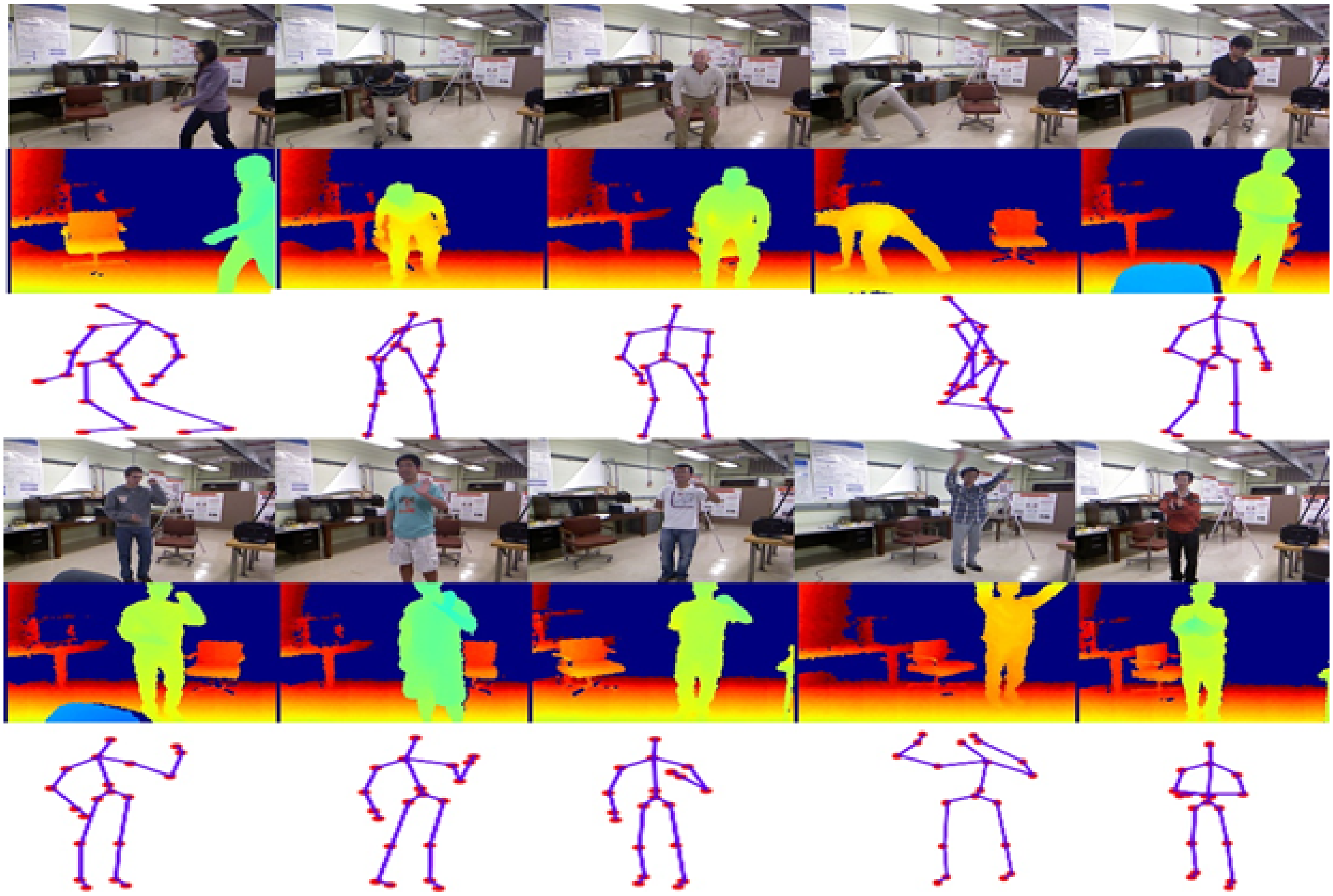


Fig. 7. Some sample frames of different activities of the UTKinectAction3D dataset in three different formats, RGB, Skeleton and Depth.

females. Each activity is captured inside the living room and there is a sofa in the scene, this means there is an interaction of subject to the object during activity capturing. Each activity is carried out by each subject once in sitting position and once in standing position. Each activity is captured in three different forms, RGB, Depth map and Skeleton joints. The file format of RGB, Depth maps and Skeleton joints are avi, bin and txt files. There are 16(activity) × 10 (person) × 2 (position) = 320 files for each modality. Therefore, 320 × 3 = 960 files in total. Fig. 6 shows a 'cheer up' activity of MSRDailyActivity3D at different time stamps in three different formats RGB, Depth and Skeleton dataset samples.

B. UTKinectAction3D Dataset

The UTKinectAction3D dataset was collected by the Microsoft Foundation Research in 2012. It contains 10 indoor actions performed by ten subjects, out of them 9 are males and 1 is female. The captured actions are 'carry', 'walk', 'sitdown', 'standup', 'push', 'pull', 'throw', 'pickup', 'wavehands', and 'claphands'. Each subject performs each activity twice. The data is captured in three formats, RGB videos, the depth map, and skeleton joints, and all activities are synchronized in all formats. The second carry activity sequence is not given hence there are total 199 activity sequences that exist in each modality. For the experimental point of view, a total of 200 sequences are used in this work. There are 10(action) × 10(subjects) × 2(frequency) = 200 activities in each case. In total 200 × 3 = 600 activities corresponding to RGB, depth, and skeleton data. The file format of the RGB video frames, depth map, and skeleton joint positions are jpg, bin, and txt file. The dataset is captured using a stationary Kinect camera. Fig. 7 shows the ten actions of UTKinectAction3D dataset.

C. Experimental Results on MSRDailyActivity3D Dataset

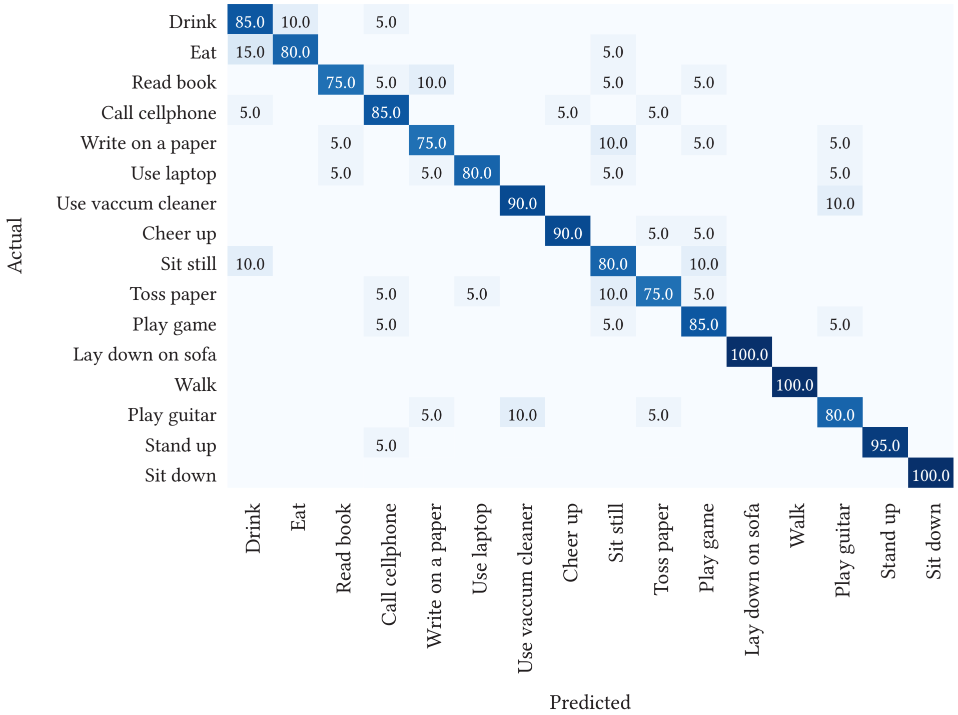
During the implementation phase, we extract the features from the RGB video sequences using a 3D Convolutional Neural Network, from the depth maps using second 3D Convolutional Neural Network and from the skeleton joints by a LSTM network for later classification by SVM. Different input cube size such as (13×50×50), (14×50×50), (15×50×50), (16×50×50) are applied to first 3DCNN and obtained the best spatio-temporal features at input size (16×50×50). Similarly for depth data, the input cubes size (13 × 32 × 32), (14 × 32 × 32), (15 × 32 × 32) and (16 × 32 × 32) have been tried to the second 3DCNN and best features are extracted at (13 × 32 × 32). To train both 3DCNN, we used a 6×10^{-4} learning rate by a decay factor that decreases with increases in the number of epochs. An Adam optimizer is used with a categorical cross-entropy loss function.

For validation of our proposed approach, we used Leave-One-User-Out cross-validation (LOUOCV). The dataset is divided into two subsets, one for the training set and another testing set. Ten folds cross-validation is used, and in each fold, nine users are used for training purpose and one user for testing. In (LOUO) cross-validation technique 9 × 16 × 2 = 288 activity sequences are used in training, and 1 × 16 × 2 = 32 activity sequences are used to perform the test in each fold. The class probability scores of each activity sequence are recorded during the testing phase.

In GA optimization, the class probability scores obtained from all three models during the cross-validation step are fused together. Here, we used the roulette wheel procedure for chromosomes selection. For the optimization, the mutation probability is set to 0.01, and the crossover rate is set to 0.8. After the successive iterations, the GA optimizes the weights and increases the classification accuracy. The GA optimization is performed with varying population sizes from 10

TABLE I. Optimized GA Results for MSRDailyActivity3D Dataset

No. of steps	No of Iteration per step	Population Size	Accuracy	w_1	w_2	w_3
1	72	10	83.44	5.429	7.367	0.854
2	51	20	84.69	9.8	9.473	6.793
3	**87**	**30**	**85.94**	**4.819**	**0.277**	**7.854**
4	76	40	83.44	2.697	3.82	4.798
5	82	50	84.37	3.339	5.197	3.763
6	68	60	84.10	2.982	-20.614	36.706
7	57	70	83.12	1.565	3.714	6.445
8	76	80	84.10	1.211	-3.296	8.612
9	67	90	83.43	1.211	-3.296	8.612
10	74	100	83.13	4.814	3.471	6.662



	Drink	Eat	Read book	Call cellphone	Write on a paper	Use laptop	Use vaccum cleaner	Cheer up	Sit still	Toss paper	Play game	Lay down on sofa	Walk	Play guitar	Stand up	Sit down
Drink	85.0	10.0		5.0												
Eat	15.0	80.0							5.0							
Read book			75.0	5.0	10.0				5.0		5.0					
Call cellphone	5.0			85.0				5.0		5.0						
Write on a paper			5.0		75.0				10.0		5.0			5.0		
Use laptop			5.0		5.0	80.0			5.0					5.0		
Use vaccum cleaner							90.0							10.0		
Cheer up								90.0		5.0	5.0					
Sit still	10.0								80.0		10.0					
Toss paper				5.0		5.0			10.0	75.0	5.0					
Play game				5.0					5.0		85.0			5.0		
Lay down on sofa												100.0				
Walk													100.0			
Play guitar					5.0		10.0			5.0				80.0		
Stand up				5.0											95.0	
Sit down																100.0

Fig. 8. Confusion Matrix of our proposed approach for MSRDailyActivity3D Dataset.

to 100. The maximum classification accuracy of **85.94%** is achieved when population size is 30 at step 3 from Table I, and corresponding optimum weight values have been recorded as w_1 = 4.819, w_2 = 0.277 and w_3 = 7.854.

It can also be observed from Table I, that performance decreases with increase in the population size, and the maximum classification accuracy is achieved at population size 30. The values of learned weight, classification accuracy, number of iterations per step along with population size, are given in Table I.

The confusion matrix of our proposed approach for the MSRDailyActivity3D dataset is given in Fig. 8. Interpretations of confusion matrix clearly show that the recognition accuracy of activities 'read book', 'write on a paper' and 'toss paper' are comparatively low due to similarity in these three activities. Likewise, the activities 'drink' and 'eat' are identical to each other due to the similar body part movement, therefore containing more confusion among them, and our model is predicting 15 percent of the times 'eat' activity as 'drink' activity. The recognition accuracy of the activities belonging to the second half of the confusion matrix shown in Fig. 8 are better than the activities belonging to the first half of the confusion matrix, since the confusion between the activities in the first half is more as compare to the activities belonging to the second half. For example, the activities 'sit down', 'walk', 'lay down on sofa' are recognized 100%, 'stand up' is also approximate to 100% except for one sample which is being misclassified as 'call cell phones', while the activities 'drink', 'eat', 'read book', 'call cell phones', 'write on a paper', 'use laptop' have lower recognition accuracies. One of the reasons for this is that, the activities such as 'read book', 'write on a paper', 'toss paper' are happening with the contact of the external object. Therefore, their recognition accuracies are less as compared to the activities in which no external object is being used.

After the above discussion, it is clear that all the activities have good recognition accuracies, except those that have higher confusion among them and those that are happening with the contact of external objects, which proves that our proposed approach learns the structural and motion information as well as relevant features from the input sequences in a better way.

In addition to applying the GA, we have also used PSO algorithm as an alternative method to optimize class scores. We repeat the optimization process ten times with unbounded weights. During the optimization process, PSO optimized the weights and improves the classification accuracy. The optimal weights and their corresponding classification accuracy are given in Table II.

TABLE II. Optimized PSO Result on MSRDailyActivity3D Dataset

No. of steps	Iteration per step	Classification Accuracy	w_1	w_2	w_3
1	31	81.56	.280	.3167	1.3071
2	30	82.19	.4702	-.9635	2.3384
3	24	83.13	.3974	.4310	1.8351
4	28	82.50	.1092	-1.5660	1.8581
5	33	82.81	2.0110	-1.1847	7.0962
6	29	82.81	0.3044	.3379	1.4145
7	27	81.87	0.3412	1.3601	2.3577
8	**30**	**83.75**	**1.0431**	**-0.2076**	**3.2864**
9	32	82.50	2.7499	1.6599	9.7201
10	24	82.19	1.8581	-.3211	5.8407

TABLE III. Performance Comparison of the Proposed Approach with Other Methods on MSRDailyActivity3D Dataset

Dataset Modality	Methods	Accuracy (%)
Depth	Only LOP feature [67]	42.5
	Histogram of oriented 4Dnormals [68]	80.0
	Depth Cuboid similarity feature (DCSF)[69]	83.6
Skeleton	NBNN [70]	53
	NBNN + time [70]	60
	NBNN + parts [70]	60
	Only Joint Position feature[67]	68
	NBNN + parts + time [70]	70
	Distance + Temporal features [71]	73.43
	mean 3D joints [72]	73.75
	SVM + FTP feature [67]	78
Hybrid	MHI+SAE feature [58]	74.6
	Actionlet Ensemble [67]	85.75
	Our Proposed Approach with GA	**85.93**
	Our Proposed Approach with PSO	**83.75**
	DCSF + joint [69]	88.2

TABLE IV. Optimized GA Results on UTKinectAction3D Dataset

No. of steps	No of Iteration per step	Population Size	Accuracy	w_1	w_2	w_3
1	47	10	95.5	7.652	14.132	10128
2	51	20	96	1.367	4.121	3.835
3	48	30	95.5	4.204	8.187	5.61
4	49	40	95.5	2.308	8.063	9.267
5	**54**	**50**	**96.5**	**1.632**	**6.529**	**8.008**
6	42	60	96	4.352	7.416	5.262
7	44	70	95.5	2.344	6.142	4.671
8	39	80	95.5	2.277	9.889	11.913
9	41	90	95	1.054	1.711	1.79
10	36	100	95.5	1.464	7.134	8.46

It can be concluded from Table II that, highest classification accuracy has been achieved in step 8. The PSO optimized weights are w_1 = 1.0431, w_2 = 0.2076 and w_3 = 3.2864, the corresponding classification accuracy is **83.75%**.

Table III contains the state-of-art methods along with their accuracy on the MSRDailyActivity3D benchmark dataset. The comparison of the recognition accuracy of our proposed method against the mentioned state-of-art methods are given in Table III and it can also be seen from Table III that the proposed approach with GA optimization has comparable accuracy with the mentioned state-of-art results.

D. Experimental Results on UTKinectAction3D Dataset

To confirm the efficacy of our method, we have evaluated our proposed approach on another benchmark UTKinectAction3D dataset. For the experiment point of view, we used the same experimental setup as used in the MSRDailyActivity3D dataset. For validation, we use leave one user out (LOUO) cross-validation in ten rounds, in which one user is removed from training and used as testing. The process is continuing for all users. In LOUO cross-validation method, 10 × 9 × 2 = 180 activities sequences are used in training, and 10 × 1 × 2 = 20 activities sequence are used in testing to test each user in each round. The probability scores of each activity sequence are recorded and fused together. Next, the GA optimizes the weights and improves the classification accuracy during successive iterations. It can be observed from Table IV, that the maximum classification accuracy 96.50% is achieved at population size=50 in step 5 from Table III, and the corresponding optimum weights are w_1 = 1.632, w_2 = 6.529 and w_3 = 8.008.

TABLE V. Optimized PSO Result on the UTKincecAction3D Dataset

No. of steps	Iteration per step	Classification Accuracy	w_1	w_2	w_3
1	42	94.5	2.3633	6.4472	-0.2244
2	**39**	**95**	**1.3515**	**3.7570**	**-1.0144**
3	37	94.5	1.4766	5.2991	0.3548
4	41	94.5	2.5806	5.5602	2.4345
5	35	94	1.5175	5.4437	0.6930
6	37	93.5	0.9930	3.4147	-0.6840
7	43	93.5	4.0628	5.8333	1.8558
8	38	94	1.6915	3.1000	-0.2983
9	41	93.5	0.3260	1.0256	0.4786
10	39	94.5	1.3927	1.8255	0.1039

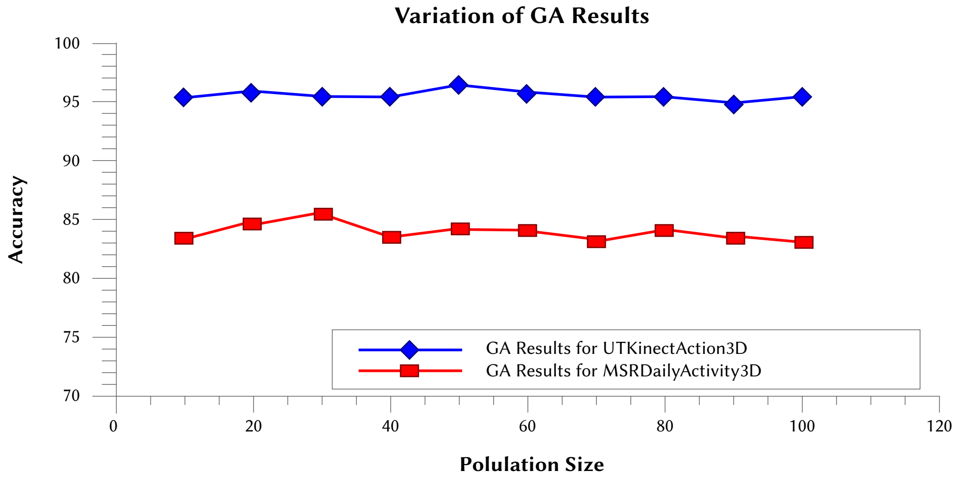


Fig. 10. Variation of GA accuracies on UTKinectAction3D and MSRDailyActivity3D datasets at different population size.

The confusion matrix of the UTKinectAction3D dataset is given in Fig. 9. It can be conclude from Fig. 9 that the activity 'throw' has comparatively less recognition accuracy due to the high degree of confusion with push activity. It can also be illustrated from the Fig. 9 that confusion occurs between the activities 'sitdown' and 'pickup' while most of the activities are 100% classified. Therefore, it proves that our proposed approach is sufficient to learn spatial and motion features from the activity sequences. The performance with the UTKinectAction3D dataset has been compared against state-of-art-methods.

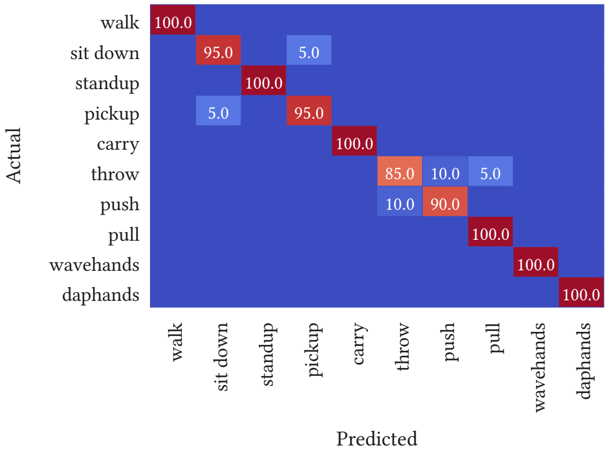


Fig. 9. Confusion matrix of UTKinectAction3D dataset.

Similarly, we have also applied the PSO algorithm on the UTKinectAction3D dataset to optimize the class scores of each test activity. The PSO optimization is performed in ten steps. The obtained results are given in the Table V.

Based on Table V, the highest classification accuracy is achieved in step 2. The PSO optimized weights are $w_1 = 1.3515$, $w_2 = 3.7570$ and $w_3 = -1.0144$ and the corresponding classification accuracy is **95.0%**.

Fig. 10 describes how the GA accuracy varies at different population size for both UTKinectAction3D and MSRDailyActivity3D datasets. The performance of UTKinectAction3D dataset has been compared against state-of-art-methods. Table VI displays the comparison results. Our proposed approach with GA optimization gives the highest classification accuracy over all other approaches.

TABLE VI. Performance of Our Proposed Approach on UTKinectAction3D Dataset, Compared to the State-of-art-approaches

Methods	Accuracy (%)
Xia. et al., (2012) [5]	90.92
Wang. et al., (2015) [49]	90.91
Liu et al., (2015) [73]	92.00
Liu. et al., (2016) [74]	96.00
Proposed Approach with GA	**96.50**
Proposed Approach with PSO	**95**

E. Error Analysis

The above discussion indicates that, during the space-time features learning using 3DCNN, we have taken the fixed number of frames in the input cube because the length of the frames in every video varies from 51 to 553 for MSRDailyActivity3D dataset and 5 to 120 for UTKinectAction3D dataset respectively, therefore some loss of information occurs.

In MSRDailyActivity3D dataset, the total false positive rate (error) is 0.141. This is because the upper half of the activities in the confusion matrix shown in Fig. 6 such as 'drink', 'eat', 'read book', 'write on a paper', 'use laptop', 'toss paper', 'sit still' and 'play guitar' have FPR 0.3125 while the remaining below half of the activities have FPR only 0.06875. The reason is that the activities in the first half of the confusion matrix are similar to each other and have more confusion compared to the activities in the second half. For example, the activities 'read book', 'write on a paper', and 'use laptop' have similar body movement with an external object. Also, the activities 'drink' and 'eat' are more confusing due to higher homogeneity levels. Furthermore, the false-positive rate (FPR) of our proposed approach for the UTKinectAction3D dataset is 0.035, out of which 0.025 error is mainly due to the two activities 'push' and 'throw'. The reason behind this is that, the uniformity between the activities 'push' and 'throw' and another reason is that the network learns almost similar motion information for both activities.

V. Conclusion and Future Work

In this paper, a Deep Multi-Model approach has been proposed for activity recognition, which mainly consists of three steps. Firstly, we extract spatial and temporal features using two 3DCNN and a LSTM networks from RGB, Depth, and skeleton information respectively. Secondly, three SVM classifiers are used to generate the class probability scores of each test activity in all formats. Finally, the class scores obtained from each modality are fused and optimized by a genetic algorithm for activity recognition. The proposed approach automatically learns high-level features from input data for each modality using spatial-temporal convolutional neural networks, and a LSTM network. Our proposed approach also uses RGB, Depth and skeleton information and gives better performance than using each modality separately. An optimization-based score fusion technique is presented to take full advantage of class label decisions from different aspects. For this purpose, we use the genetic algorithm (GA). We have evaluated our proposed approach on two benchmarks, MSRDailyActivity3D and, UTKinectAction3D datasets, and have achieved 85.94% and 96.50% recognition accuracies, respectively. The obtained results are comparable over the state-or-art-methods. Moreover, in our proposed approach, the three channels can run in parallel, therefore, it can also run on the multi-core CPU systems to save time. Hence this arrangement makes our method more efficient and fast.

Our proposed method is an application of the human-machine interface. It can be used to recognize the normal activities, abnormal activity and patient security and monitoring inside the living room etc. Despite the several advantages, our proposed system has two drawbacks. Firstly, the depth camera can capture the frames only from 4 to 11 feet. Secondly, the depth video sequences must be captured completely prior to test over the network. In future work, these challenges may be resolved.

Acknowledgment

We are grateful to the College of Engineering Roorkee, India, and UTU Dehradun, India, for providing excellent research facility to carry out this research work.

References

[1] H. Rahmani, A. Mian, and M. Shah, "Learning a deep model for human action recognition from novel viewpoints," in IEEE transactions on pattern analysis and machine intelligence, vol.40, no. 3, pp.667-681, 2017.

[2] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "Learning clip representations for skeleton-based 3d action recognition," in IEEE Transactions on Image Processing, vol. 27, no. 6, pp.2842-2855, 2018.

[3] J. K. Aggarwal, and M. S. Ryoo, "Human activity analysis: A review," in ACM Computing Surveys (CSUR), vol. 43, no. 3, pp.1-43, 2011.

[4] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3d points," in 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, San Francisco, CA, USA, 2010, pp. 9-14.

[5] L. Xia, C.C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in Computer Vision and Pattern Recognition Workshops (CVPRW), Providence, RI, USA, 2012, pp. 20-27.

[6] X. Yang,, C. Zhang, and Y. Tian, "Recognizing actions using depth motion maps-based histograms of oriented gradients," in Proceedings of the 20th ACM international conference on Multimedia, Nara, Japan, 2012, pp. 1057-1060.

[7] Y. Zhu, W. Chen, and G. Guo, "Fusing spatiotemporal features and joints for 3d action recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Portland, OR, USA, 2013, pp. 486-491.

[8] A. Chaaraoui, J. Padilla-Lopez, and F. Flórez-Revuelta, "Fusion of skeletal and silhouette-based features for human action recognition with rgb-d devices," in Proceedings of the IEEE international conference on computer vision workshops, Sydney, NSW, Australia, 2013, pp. 91-97.

[9] S. Siddiqui, M. A. Khan, K. Bashir, M. Sharif, F. Azam, and M. Y. Javed, "Human action recognition: a construction of codebook by discriminative features selection approach," in International Journal of Applied Pattern Recognition, vol. 5, no. 3, pp.206-228, 2018.

[10] A. Franco, A. Magnani, and D. Maio, "A multimodal approach for human activity recognition based on skeleton and RGB data," in Pattern Recognition Letters, vol. 131, pp. 293-299, 2020.

[11] K. Khoshelham, and S. O. Elberink, "Accuracy and resolution of kinect depth data for indoor mapping applications," in Sensors, vol. 12, no. 2, pp. 1437-1454, 2012.

[12] H. H. Pham, L. Khoudour, A. Crouzil, P. Zegers and S. A. Velastin, "Exploiting deep residual networks for human action recognition from skeletal data," in Comput Vis Image Underst, vol. 170, pp. 51–66, 2018.

[13] S. Yang, J. Yang, F. Li, G. Fan and D. Li, "Human Action Recognition Based on Fusion Features," in International Conference on Cyber Security Intelligence and Analytics, 2019, pp. 569–579.

[14] A. Jalal, M. Z. Uddin, and T. S. Kim, "Depth video-based human activity recognition system using translation and scaling invariant features for life logging at smart home," in IEEE Transactions on Consumer Electronics, vol. 58, no. 3, pp.863-871, 2012.

[15] M. Khan, T. Akram, M. Sharif, N. Muhammad, M. Javed and S. Naqvi, "An improved strategy for human action recognition; experiencing a cascaded design," in IET Image Processing, vol 14, no. 5, pp. 818-829, 2019.

[16] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition", in IEEE conference on computer vision and pattern recognition, Los Vegas, NV, USA, 2016, pp. 770–778.

[17] L. Bi, D. Feng and J. Kim, "Dual-path adversarial learning for fully convolutional network (FCN)-based medical image segmentation," in Visual Computers, vol. 34, no. 6, pp. 1–10, 2018.

[18] M. Rashid, M. A. Khan, M. Sharif, M. Raza, M. M. Sarfraz and F. Afza, "Object detection and classification: a joint selection and fusion strategy of deep convolutional neural network and SIFT point features," in Multimedia Tools and Applications, vol. 78, no. 12, pp.15751–15777, 2019.

[19] F. Zhou, Y. Hu and X. Shen, "Msanet: multimodal self-augmentation and adversarial network for RGB-D object recognition," The Visual Computers, vol. 35, no. 11, pp. 1583-1594, 2019, https://doi.org/10.1007/s00371-018-1559-x

[20] I. Gogić,, M. Manhart, I. S. Pandžić and J. Ahlberg,, "Fast facial expression recognition using local binary features and shallow neural networks," in The Visual Computer, vol. 36, no. 01, pp.1-16, 2018.

[21] M. Sharif, M. A. Khan, M. Faisal, M. Yasmin and S. L. Fernandes, "A framework for offline signature verification system: Best features selection approach," in Pattern Recognition Letters, 2018.

[22] K. K. Verma, B. M. Singh and A. Dixit, "A review of supervised and unsupervised machine learning techniques for suspicious behavior recognition in intelligent surveillance system", in International Journal of Information Technology, 2019, pp. 1-14.

[23] G. I. Parisi, "Human Action Recognition and Assessment via Deep Neural Network Self-Organization," in Modelling Human Motion, pp. 187-211, 2020.

[24] X. X. Niu and C. Y. Suen, "A novel hybrid CNN–SVM classifier for recognizing handwritten digits," in Pattern Recognition, vol. 45, no. 4, pp. 1318-1325, 2012.

[25] D. X. Xue, R. Zhang, H. Feng and Y. L. Wang, "CNN-SVM for microvascular morphological type recognition with data augmentation," in Journal of medical and biological engineering, vol. 36, no. 6, pp. 755-764, 2016.

[26] A. B. Sargano, X. Wang, P. Angelov and Z. Habib, "Human action recognition using transfer learning with deep representations," in 2017 International joint conference on neural networks (IJCNN), Anchorage, AK, USA, 2017, pp. 463-469.

[27] T. Jiang, Z. Zhang and Y. Yang, "Modeling coverage with semantic embedding for image caption generation," in The Visual Computers, vol. 35, no. 11, pp. 1655-1665, https: //doi.org/10.1007/s00371-018-1565-z

[28] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei and S. Savarese, "Social lstm: human trajectory prediction in crowded spaces," in IEEE conference on computer vision and pattern recognition, 2016, pp 961–971

[29] I. Sutskever, O. Vinyals Q. V. Le, "Sequence to sequence learning with neural networks," in Advances in neural information processing systems, 2014, pp. 3104–3112.

[30] K.K. Verma, B. M. Singh, "Deep Learning Approach to Recognize COVID-19, SARS and Streptococcus Disease from Chest X-Ray Images," in Journal of Scientific and Industrial Research, vol. 80, no. 01, pp. 51-59, 2021.

[31] J. Cong and B. Zhang, "Multi-model feature fusion for human action recognition towards sport sceneries," in Signal Processing: Image Communication, 2020.

[32] E. Zhou and H. Zhang, "Human action recognition towards massive-scale sport sceneries based on deep multi-model feature fusion," Signal Processing: Image Communication, vol. 84, 2020.

[33] K. Soomro, A. R. Zamir and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," arXiv preprint arXiv:1212.0402, 2012.

[34] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio and T. Serre, "HMDB: a large video database for human motion recognition," in 2011 International Conference on Computer Vision, Barcelona, Spain, 2011, pp. 2556-2563.

[35] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in Proceedings of the IEEE conference on computer vision and pattern recognition, Boston, MA, USA, 2015, pp. 961-970.

[36] A. B. Sargano, P. Angelov and Z. Habib, "A comprehensive review on handcrafted and learning-based action representation approaches for human activity recognition," in Applied sciences, vol. 7, no. 01, 2017.

[37] H. Wang, and C. Schmid, "Action recognition with improved trajectories," in Proceedings of the IEEE international conference on computer vision, Sydney, NSW, Australia, 2013, pp. 3551-3558.

[38] A. Krizhevsky, I. Sutskever and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, vol. 25, pp. 1097-1105, 2012.

[39] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in Advances in neural information processing systems, 2014, pp. 568-576.

[40] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014, pp. 1725-1732.

[41] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in Proceedings of the IEEE international conference on computer vision, Santiago, Chile, 2015, pp. 4489-4497.

[42] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, NV, USA, 2016, pp. 1933-1941.

[43] G. Varol, I. Laptev and C. Schmid, "Long-term temporal convolutions for action recognition," in IEEE transactions on pattern analysis and machine intelligence, vol. 40, no. 6, pp.1510-1517, 2017.

[44] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad and S. W. Baik, "Action recognition in video sequences using deep bi-directional LSTM with CNN features," in IEEE Access, vol. 6, pp.1155-1166, 2017.

[45] K. K. Verma, B. M. Singh, H. L. Mandoria and P. Chauhan, "Two-Stage Human Activity Recognition Using 2D-ConvNet," in International Journal of Interactive Multimedia & Artificial Intelligence, vol. 6, no 2, pp. 135-135, 2020.

[46] Z. Li, Z. Zheng, F. Lin, H. Leung and Q. Li, "Action recognition from depth sequence using depth motion maps-based local ternary patterns and CNN," in Multimedia Tools and Applications, vol. 78, no. 14, pp.19587-19601, 2019.

[47] P. Wang, W. Li, Z. Gao, C. Tang and P. O. Ogunbona, "Depth pooling based large-scale 3-d action recognition with convolutional neural networks," IEEE Transactions on Multimedia, vol. 20, no. 5, pp.1051-1061, 2018.

[48] C. Chen, R. Jafari and N. Kehtarnavaz, "Action recognition from depth sequences using depth motion maps-based local binary patterns," in 2015 IEEE Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 2015, pp. 1092-1099.

[49] P. Wang, W. Li, Z. Gao, J. Zhang, C. Tang and P. Ogunbona, "Deep convolutional neural networks for action recognition using depth map sequences," arXiv preprint arXiv:1501.04686, 2015.

[50] V. Megavannan, B. Agarwal and R. V. Babu, "Human action recognition using depth maps," in 2012 International Conference on Signal Processing and Communications(SPCOM), Bangalore, India, 2012, pp. 1-5.

[51] Y. Han, S. L.Chung, Q. Xiao, W. Y. Lin and S. F. Su, "Global Spatio-Temporal Attention for Action Recognition based on 3D Human Skeleton Data," in IEEE Access, vol. 8, pp. 88604-88616, 2020.

[52] B. Ren, M. Liu, R. Ding and H. Liu, "A Survey on 3D Skeleton-Based Action Recognition Using Learning Method," arXiv preprint arXiv:2002.05907, 2020.

[53] R. Saini, P. Kumar, P. P. Roy and D. P. Dogra, "A novel framework of continuous human-activity recognition using kinect," in Neurocomputing, vol. 311, pp.99-111, 2018.

[54] B. Chikhaoui, B. Ye and A. Mihailidis, "Feature-level combination of skeleton joints and body parts for accurate aggressive and agitated behavior recognition," in Journal of Ambient Intelligence and Humanized Computing, vol. 8, no. 6, pp.957-976, 2017.

[55] A. Shahroudy, J. Liu, T.T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, NV, USA, 2016, pp. 1010-1019.

[56] Y.Gu, X Ye, W. Sheng, Y. Ou and Y. Li, "Multiple stream deep learning model for human action recognition," in Image and Vision Computing, vol. 93, 2020.

[57] P. Khaire, P. Kumar and J. Imran, "Combining CNN streams of RGB-D and skeletal data for human activity recognition," in Pattern Recognition Letters, vol. 115, pp.107-116, 2018.

[58] A. Tomas and K. K. Biswas, "Human activity recognition using combined deep architectures," in 2017 IEEE 2nd International Conference on Signal and Image Processing (ICSIP), Singapore, 2017, pp. 41-45.

[59] E. P. Ijjina and K. M. Chalavadi, "Human action recognition in RGB-D videos using motion sequence information and deep learning," in Pattern Recognition, vol. 72, pp. 504-516, 2017.

[60] C Zhao, M. Chen, J. Zhao, Q. Wang and Y. Shen, "3D Behavior Recognition Based on Multi-Modal Deep Space-Time Learning," in Applied Sciences, vol. 9, no. 4, pp.716, 2019.

[61] S. Ji, W. Xu, M. Yang and K. Yu, "3D convolutional neural networks for human action recognition," in IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 01, pp. 221-231, 2012.

[62] S. Hochreiter and J. Schmidhuber, "Bridging long time lags by weight guessing and "Long Short-Term Memory"," in Spatiotemporal models in biological and artificial systems, vol. 37, pp. 65-72, 1996.

[63] S. Gaglio, G. L. Re and M. Morana, "Human activity recognition process using 3-D posture data," in IEEE Transactions on Human-Machine Systems, vol. 45, no. 05, pp.586-597, 2014.

[64] O. Chapelle, V. Vapnik, O. Bousquet and S. Mukherjee S, "Choosing

multiple parameters for support vector machines," in Machine learning, vol. 46, no. 1-3, pp. 131-59, 2002.

[65] D. E. Goldberg, B. Korb and K. Deb, "Messy genetic algorithms: Motivation, analysis, and first results," in Complex systems, vol. 3, no. 05, pp.493-530, 1989.

[66] J. Kennedy and R. Eberhart, "Particle swarm optimization," in Proceedings of ICNN'95-International Conference on Neural Networks, Perth, Australia, 1995, pp. 1942-1948.

[67] J. Wang, Z. Liu, Y.Wu and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 2012, pp. 1290-1297.

[68] O. Oreifej and Z. Liu, "Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences," in Proceedings of the IEEE conference on computer vision and pattern recognition, Portland, OR, USA, 2013, pp. 716-723.

[69] L. Xia and J. K. Aggarwal, "Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera," in Proceedings of the IEEE conference on computer vision and pattern recognition, Portland, OR, USA, 2013, pp. 2834-2841.

[70] L. Seidenari, V. Varano, S. Berretti, A. Bimbo and P. Pala, "Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Portland OR, USA, 2013, pp. 479-485.

[71] Y. Hbali, S. Hbali, L. Ballihi and M. Sadgal, "Skeleton-based human activity recognition for elderly monitoring systems," in IET Computer Vision, vol. 12, no. 01, pp.16-26, 2017.

[72] A. Ben Tamou, L. Ballihi and D. Aboutajdine, "Automatic learning of articulated skeletons based on mean of 3d joints for efficient action recognition," in International Journal of Pattern Recognition and Artificial Intelligence, vol. 31, no. 04, 2017.

[73] A. A. Liu, W. Z. Nie, Y. T. Su, L. Ma, T. Hao and Z. X. Yang, "Coupled hidden conditional random fields for RGB-D human action recognition," in Signal Processing, vol. 112, pp. 74-82, 2015.

[74] Z. Liu, C. Zhang, Y. Tian, "3D-based deep convolutional neural network for action recognition with depth sequences," in Image Vis. Comput., vol. 55, pp. 93–100, 2015.

Kamal Kant Verma

Kamal Kant Verma is research scholar in Uttarakhand Technical University Dehradun. He is currently working as an Assistant Professor in Department of Computer Science & Engineering, COER Roorkee Uttarakhand India. He did B.Tech in Information Technology in 2006, M.Tech in CSE in 2012 and currently pursuing PhD from Uttarakhand Technical University Dehradun India. He has 15 years of teaching and research experience. His research area is Human Activity Recognition, Human Computer Interface, Pattern Recognition and Signal Processing. He has published more than 20 research papers in reputed national/ international journal and conferences such as Springer, Elsevier, IJIMAI, etc.

Brij Mohan Singh

Brij Mohan Singh is Director of College of Engineering Roorkee & Professor in Department of Computer Science and Engineering, COER Roorkee India. He has published more than 35 research papers in International Journals such as Document Analysis and Recognition-Springer, CSI Transactions on ICT-Springer, IJIG-World Scientific, IJMECS, EURASIP Journal on Image and Video Processing etc. His research areas are DIP and Pattern Recognition. He has guided 3 PhD Thesis of UTU and currently 6 are in process.